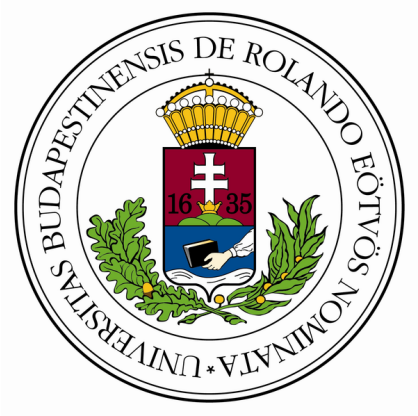
Protein structure and evolution in the light of flexibility

- Doctoral Thesis Statements -



Annamária Franciska Ángyán

Doctoral School of Chemistry
Head of Doctoral School: Dr. György Inzelt, D.Sc.
Synthetic Chemistry, Materials Science and Biomolecular Chemistry Doctoral Program
Head of Program: Dr. András Perczel, D.Sc.

Supervisor: Dr. Zoltán Gáspári, Ph.D.

Eötvös Loránd University
Institute of Chemistry
Laboratory of Structural Chemistry and Biology

Budapest
- 2012 -

Introduction

In the last years, molecular biology witnessed a number of paradigm changes. The internal dynamics of proteins came into focus and under protein structure, we now mean not only a single globular state, but other conformations of the polypeptide chain. The extreme case of this is represented by the so-called intrinsically disordered proteins.

During the protein-ligand binding process, both the structure and dynamics of the protein changes. This is connected to the the entropic part of the binding free energy. The "folding and binding" mechanism is spectacular for intrinsically disordered proteins, and is observable for globular proteins too. Numerous publications underly the importance of conformational selection, but only few studies present it at atomic level. NMR spectroscopy is the most appropriate method for the description of the structure and dynamics of proteins. The experiments are made under nearly physiological-like conditions, in solution. The NMR sample contains approximately 10^{16}-10^{17} protein molecules. The flexibility and conformational heterogeneity of proteins in solution cannot be described by a single conformer. All measured parameters represent and average over time and across the ensemble, so a single-conformer model can be inappropriate and insufficient. There are methods capable of calculating conformational ensembles reflecting the internal dynamics of the studied protein at different timescales. A growing number of publications recognizes the importance of conformational selection, but only few atomic levels are available to describe this processes in detail.

Various processes, such as ligand binding, mutations or disorder-to-order transitions can induce changes in protein dynamics. At evolutionary scale, protein conformational diversity play a crucial role and protein dynamics has a strong effect on its biological function. The emerging consensus on protein aggregation is that it is an inherent property of any polypeptide chain and, regardless of their amino acid sequences, the amyloid fibril might be the most favored thermodynamic state of all proteins. Even so, proteins display sequence-specific aggregation propensities that can be estimated by *in silico* methods. Thus, proteins can evolve to reduce the risk of aggregation and detailed studies of selected proteins revealed a number of such mechanisms. However, proteins continuously emerge *de novo* by multiple mechanisms, including the transcription and translation of previously non-coding DNA segments. This poses the question whether novel proteins that did not yet have the chance to reduce their aggregation load by selection can seriously hinder molecular evolution: if the aggregation propensity of *de novo* proteins is generally high, leading to the aggregation of practically all *de novo* polypeptides, that might render the chances of the emergence of such proteins negligible.

Protein dynamics and evolution will be studied in the light of flexibility. In the first part, I present new methods related to dynamic structural ensembles and the assessment

of structure accuracy. This part is based on the first three publications. The second part places in evolutionary scale the dynamic properties of *de novo* proteins. This part is based on the fourth publication.

First part: Assessment of NMR experimental data with structural ensembles

Aims

In the first part, I present a fast method capable to relate known folds to the obtained NOE data, and an other method for evaluating the accuracy of dynamic conformational ensembles.

Methods

- I used the RCSB PDB, the RECOORD, and the SCOP ASTRAL databases. The protein structures were downloaded from the PDB and RECOORD databases . The hierarchical SCOP database was used during the development.
- The molecular dynamic simulations were done using the GROMACS program, with the OPLS-AA force field. The SHIFTX and PALES programs were implemented.
- The method development and data analysis was done with own PERL programs. The servers are written in PERL and C++ program language. The statistical analysis was done with contingency-analysis.

Results

1. I developed the PRIDE-NMR algorithm, suitable for evaluating protein folds based on the H-H distance distributions. I tested the method on more datasets (40-protein test set, PDB dataset). The test set composition was similar to those used by Novotny for evaluating the protein fold comparison servers. The efficiency of the PRIDE-NMR method is in the same range as those for the best protein fold estimation methods.

2. The PRIDE-NMR method is available as a web server. The background database cutoff, the minimal distance bin taken into account, the chain length weighting and the percentage weighting can be adjusted in order to give a wide range of results. I presented the use of the server on the example of an SH3 domain.

3. I truncated the back-calculated datasets and the NOE data from 10 to 85%, by 15% steps. The results obtained show that the back-calculated H-H pairs are unable to characterize specifically the fold. The 10 times sparser NOE data, even truncated, are able to find similar folds.

4. I analyzed the correspondence between three dynamic structural ensembles (DER, MUMO and a new MUMO simulation) for ubiquitin with a distance restrains list containing only unambiguous restraints. The results show that the PRIDE-NMR method is suitable for quality evaluation of dynamic structural ensembles too.

5. We constructed the web application CoNSEnsX (Consistency of NMR-derived Structural Ensembles with eXperimental data) allowing fast, simple and convenient assessment of the correspondence of the ensemble as a whole with diverse independent NMR parameters available. The CoNSEnsX application evaluates NOEs and NMR parameters reflecting the internal dynamics of the protein such as RDCs, S^2, J-couplings, chemical shifts. Our results present a new conceptual method for the evaluation of dynamic conformational ensembles resulting from NMR structure determination. The designed CoNSEnsX approach gives a complete evaluation of these ensembles and is freely available as a web service. We have chosen different ensembles of proteins, human ubiquitin, and a disordered subunit of cGMP phosphodiesterase 5/6 for detailed evaluation and demonstration of the capabilities of the CoNSEnsX approach.

Conclusions

- The PRIDE-NMR method evaluates the protein fold using distance restraints. I showed that the NOE data contain sufficient specific information to give a first guess about the expected protein structure.

- With NMR parameters, we can describe the structure and internal dynamics and flexibility of the molecule. Due to the fact that experiments are made in solution, the results obtained are expected to be more in accordance with the real biological system.

- By taking into account the molecular motions on different timescales during the molecular dynamics simulation, the conformational heterogeneity of the resulting structural ensemble will be in accordance with the internal dynamics of the protein at a given timescale and not resulting from the distance restraints uncertainity.

Second part: *In silico* structural study of random-sequence proteins

Aims

In the second part, the protein flexibility is evaluated on an evolutionary scale. I address the intrinsic aggregation propensity of hypothetical proteins emerging *de novo* by transcription and translation of previously non-coding genomic sequences, a scenario that has been recently shown to be responsible for the evolutionary appearance of several human proteins. I investigated whether such sequences pose an 'aggregation threat' to the organisms, i.e. whether they have a higher aggregation potential than extant sequences with considerable selection against aggregation behind them.

Methods

- I generated 10,000 random DNA sequences of 480 nucleotides without in-frame STOP codons for each of GC-content regime from 10% to 90% using steps of 10%. The 160-residue length of the translated polypeptides can be regarded as a reasonable estimate of average domain size in proteins. After translating all of the 9 x 10,000 nucleotide sequences, we have used BLAST search to assess the similarity of the resulting random de novo proteins to known sequences.

- The human and mouse proteomes were randomized in order to obtain an other type of random sequences. The number of sequences, the sequence lengths and the amino-acid composition of each sequence were maintained. Each sequence was shuffled N times, where N is the number of residues in the given sequence

- I included in the study several databases representing folded, disordered, transmembrane and aggregation-prone proteins as well as the complete human and mouse proteomes. These databases were homology-filtered before use.

- The coding sequences of human orphan proteins were obtained by comparing the translated mRNA sequences obtained from the UniProt database to the available protein sequences and extracting the nucleotide sequences in the matching region.
- I used a set of prediction algorithms to assess their aggregation loads, their disorder and transmembrane propensities. None of the applied methods uses evolutionary information during data processing.
- Calculation of Pearson correlation coefficients as well as one- and two-dimensional Kolmogorov-Smirnov tests were performed. The overlapping areas were calculated by laying a grid to the 2D point distributions with a resolution of 0.1.

Results

1. According to the averaged structural predictions, the GC-content of the underlying DNA sequences governs the structural preferences of the random proteins with clearly identifiable trends that are much more pronounced than the variations in the simple physico-chemical parameters. Intrinsic disorder is a dominant feature of sequences with coding regions of high GC-content. The propensity to form transmembrane helices is relatively high at low GC-content and decreases rapidly, similarly to the aggregation load of random proteins.
2. Statistical tests reveal that the distributions obtained for the disorder, transmembrane and aggregation tendencies of the human proteome and the proteins translated from random DNA segments with 40-50-60% GC content are totally dissimilar with a P-value of 0. This is due to the different local densities of the data points in he investigated data sets. However, when estimating the (2D or 3D) space covered by the data points corresponding to the random sequence set above, it is apparent that more than 95% of this space overlaps with that spanned by proteins in the human proteome.
3. The trends observed for the random-sequence proteins are not chain length dependent. I observed the same trend by comparing the human and the random human proteome as by comparing the human and the 160 residue random-sequence proteins.
4. I investigated the three *de novo* human proteins using the same methodology as for random sequences. Interestingly, these are in accordance with the trends observed

for random de novo proteins with respect to the dependence of structural features on the GC-content of the underlying DNA segment.

Conclusions

- My study corresponds to a first approximation of the problem and can rather be viewed as a benchmark study than an accurate model of real evolutionary processes.
- The random proteins translated from DNA with 40-60% GC occupy a region in the space of the properties considered that is almost entirely within the span of those of extant proteins in the human proteome. Random de novo proteins are not expected to have a larger aggregation potential than existing ones, nor display a higher degree of disorder. However, they clearly display a lower propensity to form transmembrane helices, meaning that from the three properties investigated, this is the one that most likely needs the most serious optimization during further evolution.
- My results do not contradict the presence and nature of selection pressures present at any later stages of protein evolution, but they suggest that the appearance of novel coding sequences is not expected to be hampered by unusually high aggregation propensity of the translated proteins.

Summary

In my work I investigated dynamic and evolutionary aspects of proteins The PRIDE-NMR method is a fast novel method capable to relate known protein folds using NMR distance restraints. CoNSEnsX is a conceptually new method for the evaluation of dynamic conformational ensembles resulting from NMR structure determination. My finding that de novo proteins are not particularly prone to aggregation might appear contradictory to claims that proteins are optimized against aggregation during evolution. However, our methods addressing three basic structural properties do not reveal any detailed structural, let alone functional features. It is expected that after the birth of a de novo protein it is optimized by selection to perform its function and to adjust its structure, stability and dynamics. During this process the maintenance or even lowering of the aggregation potential present in the newly born protein is one of the pressures operative during evolution.

In summary, during my work, I developed two servers assessing the compliance of NMR parameters with structural ensembles and a new kind of *in silico* structural study evaluating the structural preferences of random *de novo* protein sequences. The protein structure and evolution has been evaluated in the light of flexibility.

Publications on the subject of the dissertation

I. **Ángyán AF**, Perczel A, Pongor S, Gáspári Z:
Fast protein fold estimation from NMR-derived distance restraints.
Bioinformatics 24:272-275 **(2008)**
impact factor: 4.328

II. Gáspári Z, **Ángyán AF**, Dhir S, Franklin D, Perczel A, Pintar A, Pongor S:
Probing dynamic protein ensembles with atomic proximity measures.
Current Protein & Peptide Science 11:515-522 **(2010)**
impact factor: 3.830

III. **Ángyán AF**, Szappanos B, Perczel A, Gáspári Z:
CoNSEnsX: an ensemble view of protein structures and NMR-derived experimental data.
BMC Structural Biology 10:39 **(2010)**
impact factor: 2.258

IV. **Ángyán AF**, Perczel A, Gáspári Z:
Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: is aggregation the main bottleneck?
FEBS Letters, before acceptance (minor revision)

Conference posters and lectures

1. **2006.** 3^{rd} Central European Conference on Chemistry towards Biology (CTB3), Krakkó, Lengyelország, poszter címe: *Fast protein fold estimation based on NMR distance restraints.* **Ángyán AF**, Perczel A, Gáspári Z.
2. **2006.** XII. Vegyészkonferencia, Csíkszereda, Románia, előadás címe: *Fehérje-térszerkezetek gyors becslése NMR kényszerfeltételek alapján.* **Ángyán AF**, Perczel A, Gáspári Z.
3. **2008.** 4^{th} Central European Conference on Chemistry towards Biology (CTB4), Dobogókő, Magyarország, poszter címe: *Fast protein fold estimation based on NMR distance restraints derived from NMR spectra (Identifying protein fold from assigned NMR spectra)* **Ángyán AF**, Perczel A, Pongor S, Gáspári Z.
4. **2009.** ChemAxon User Group Meeting Training Day, Budapest, poszter címe: *Evaluation of small molecular libraries using molecular docking and binding profile analysis.* **Ángyán AF**, Iván G, Grolmusz V.
5. **2009.** 17^{th} Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) 8th European Conference on Computational Biology (ECCB), Stockholm, Svédország, poszter címe: *Evaluating small molecule libraries using molecular docking and binding profile analysis.* **Ángyán AF**, Iván G, Grolmusz V.
6. **2009.** Foldamers: building blocks, structure and function, Szeged, poszter címe: *Generation and analysis of realistic protein structural ensembles (What do dynamic protein structural ensembles tell us?)* Gáspári Z, **Ángyán AF**, Szappanos B, Várnai P, Pongor S, Perczel A.
7. **2010.** FEBS Congress, Göteborg, Svédország, előadás és poszter címe: *CoNSEnsX: an ensemble view of protein and peptide structures ans NMR-derived experimental data.* **Ángyán AF**, B Szappanos B, Perczel A, Z Gáspári Z.